DOCUMENT RESUME

ED 287 288 FL 016 944

AUTHOR Odlin, Terence
TITLE Some Problems Concerning the Interpretation of Passage Correction Tests.
PUB DATE 87
NOTE 9p.; In: Language Testing Research; Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986); see FL 016 938.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Comparative Analysis; *Discourse Analysis; *English (Second Language); Error Patterns; Higher Education; Holistic Evaluation; *Language Tests; Scoring; Second Language Instruction; *Test Format; Testing Problems; Test Interpretation; Visual Stimuli; *Writing Evaluation; Writing Skills
IDENTIFIERS *Error Correction (Language)

ABSTRACT

A study investigated the problem of systematic oversights by students taking language passage correction (PC) tests. In these tests, students are asked to correct errors inserted in written prose passages. When errors are not overtly marked, students often overlook them. This study examined patterns of oversight in two PC tests used to evaluate skills in English as a second language (ESL). In one test, students had to consult a picture to correct semantic inaccuracies and other errors. In the other, students had to correct a written passage without consulting a picture. Results of the picture test were compared with scores on a standardized English test, results of the second test were compared with holistic evaluations of class essays, and for both tests, the results obtained from native and non-native English speakers were compared. Results of some of the analyses suggest that PC tests may be useful for research in second language learning and for second language instruction. However, some patterns in the test results suggest potential problems in interpretation. These include differences among native speakers' performances, discrepancies between PC and holistic evaluation results, and low inter-item correlations despite a high internal consistency measurement. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED287288

Some Problems Concerning the Interpretation of Passage Correction Tests

Terence Odlin
Ohio State University

Passage Correction tests (PC) are measures which require individuals to identify and correct errors that have been inserted in a prose passage. There is a family of editing test formats that could fit the above definition, but a distinctive characteristic of the tests to be described herein is that no errors are overtly marked. When individuals have no warning about what in the passage is anomalous, they may and often do overlook errors. Such systematic oversights can be a valuable source of data for second language acquisition researchers and for teachers of writing.

Earlier studies of PCs by Davies (1975), Bowen (1978), Arthur (1980), and others have shown that PCs have some interesting similarities and differences with other language tests. The similarity of passage correction to the editing of one's own writing makes the PC attractive for programs teaching basic writing, and university researchers in England, New Zealand, and the United States have either used or have considered using PCs in their ESL programs. Nevertheless, passage correction tests have not attracted nearly as much attention as cloze tests or other measures, and so it is natural that PCs are much less understood. PCs deserve to be better understood, however, since they have considerable potential to produce interesting data on language transfer, monitoring, and the development of basic writing skills. Accordingly, this paper deals with results from two PCs that I have been studying over the last five years. The results suggest problems of interpretation that may arise with a wide variety of PCs.

The two tests to be described involve essentially the same task: students in ESL courses had to detect different types of errors and correct them. However, in other details the two measures differ considerably. One of the PCs, hereafter referred to as the Picture Test, was about two hundred words long and was accompanied by a picture. Test takers had to consult the picture in order to correct semantic inaccuracies as well as other types of errors (Odlin 1986). The other test, hereafter referred to as the Fossil Test, was about a thousand words long and was not accompanied by any picture. The text of the Fossil Test was a description of fossils that was written for a non-specialist audience having little training in science (cf. Odlin 1985). The Fossil Test was considerably more difficult than the Picture Test but, as the discussion will show, it has proven useful as a measure of performance in basic writing courses. Three issues relevant to one or the other (or both) of the tests will be discussed in this paper: 1) the use of native speakers of English as a baseline for gauging the performance of ESL students; 2) the comparative results obtained on the Fossil Test and on holistically graded essays; and 3) some results of item analyses.

FL016944

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Schoffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Before these three issues are discussed, it will help to identify some general results from the two PCs. In a number of respects, the track records of both tests are good. The reliability coefficients (KR-20 and test-retest) have generally been between .80 and .90, a fact which is especially significant in view of the small number of items on both tests (26 on the Picture Test and 24 on the Fossil Test). Another characteristic of both measures is that the results correspond to results from other types of placement tests. In the case of the Picture Test, the results were compared with those on the Comprehensive English Language Test (CELT) and there were correlations between .83 and .87 ($p<.01$ in all cases) when different scoring systems were used. The Fossil Test results, which will be discussed in detail below, show a stratification similar to that obtained on a placement exam involving essays scored holistically. Moreover, both tests show that native speaker control groups outperformed ESL students.

In contrast to these similarities between the two PCs, there are test-specific characteristics that are noteworthy. The Picture Test showed significant correlations between sub-types of items that were semantically or grammatically related in different ways, and results suggested that individuals sensitive to certain types of semantic anomalies were likely to score high on the CELT. The Fossil Test indicated that students who received special training in error detection showed a significant score improvement while individuals who received no training showed no improvement. Such results suggest the potential of the PC format not only for further research in second language acquisition but for second language teaching as well. Using PCs well presupposes understanding them well, however, and it is thus desirable to consider now some of the factors related to interpretation of such tests.

The Native Speaker Baseline

The results from the Picture Test indicate a clear difference in native and non-native abilities. Table 1 shows that the control group of 20 native speakers had little difficulty with the test whereas the ESL students generally had great difficulty, no matter what their ability level was. While the native speaker scores are not perfect, the mean is nearly twice as high as that of the most advanced of the ESL groups. This result, along with the fact that there was no consistent pattern of items missed by the native speakers, makes it reasonable to attribute the less-than-perfect scores of the control group to oversights on their part.[1]

Table 1

Items Correct on Picture Test

	IEP 0-1	IEP 2-3	IEP 4-5	IEP 0-5	CONTROL
Mean	1.5	6.3	13.8	7.2	23.1
S.D.	2.4	3.7	2.8	5.8	2.0
N	8	9	8	25	20

[1] The less than perfect scores cannot be attributed to defective test items. Such items were excluded from the scoring (for details, cf. Odlin 1986).

In Table 1 IEP stands for Intensive English Program. The maximum possible score was 26.

The Fossil Test results also showed differences between native and non-native speakers, but there are some interesting wrinkles in these results. First, two different native-speaker groups that have taken the test have produced baseline scores that are significantly different. One group consisted of 30 individuals from Ohio State University who were enrolled in an upper-division class on traditional English grammar; the second group consisted of 35 freshmen at the University of Texas at Austin who were enrolled in two different sections of an English composition course. The composition course was not, technically, a remedial writing course. However, many of the students in the course had been only conditionally admitted to the University, and their instructors confirmed what I had heard from other instructors teaching the same course: namely, that the provisionally admitted students were considerably weaker in English than were freshmen who had been admitted unconditionally. In contrast, the English grammar class consisted largely of juniors and seniors with some graduate students, and in some cases these individuals were English teachers or teacher-trainees. Not too surprisingly, then, the students in the grammar course scored higher than those in the composition course, with the scores being significantly different on a t-test, as shown in Table 2.

Table 2

Native-speaker Scores on Fossil Test

	AGC	FCC
Mean	17.0	13.2
S.D.	3.0	2.9
N	30	35

Maximum possible score = 24; $t = 5.63$, $p < .01$ (two-tailed)

FCC = Freshmen composition class; AGC = Advanced grammar class

These results led to a number of questions:

1) Could the differences in native-speaker performance be due to some difference in linguistic proficiency? The answer most probably is "No." Only individuals who were native speakers of English were included in the scoring, and there are no indications that regional dialects played any role in the results.

2) Are differences in metalinguistic awareness involved? Here, the answer most probably is "Yes." The differences between the two native speaker groups already described strongly suggest that the students in the grammar course had a wider-ranging awareness of linguistic forms and functions.

3) Do the less-than-perfect scores result from inappropriate items? Here, the answer most probably is "No." When individuals failed to notice an error,

the reason could generally be attributed to the location of the errors--either the location within sentences or within paragraphs. More details on the locations of errors will appear below.

4) Did individuals have enough time to complete the test? Here, the answer is less certain. For the sake of uniformity in testing conditions, all groups taking the Fossil Test have been limited to 30 minutes. It is likely that extra time would have improved some scores. However, 30 minutes did prove to be sufficient for some individuals to obtain near-perfect scores. In addition, many native speakers succeeded in identifying 24 real or imagined errors in the text, and so extra time would probably not result in significant changes in the scores of such individuals.[2]

Comparative Results of PCs and Holistic Evaluation

The results to be discussed in this section come from a comparison of student performance on the Fossil Test with student performance on essays scored holistically by staff in the ESL writing program at Ohio State University. While the PC results generally support the validity of holistic evaluation, they indicate some potential problems with this scoring procedure.

The writing program consists of three courses, 106, 107, and 108, entry into which is initially determined by an essay examination that is scored holistically.[3] All exams are read by two program staff members, who invariably have had considerable training and experience with holistic grading, and in cases where their placement recommendations do not coincide, a third reader is consulted. For foreign students, admission to Ohio State requires a minimum TOEFL score of 500, and most of the students taking the essay exam have TOEFL scores between 500 and 600. Students placed into a 106 course normally take 107 and 108 later the same year, and their progress in each course is judged as satisfactory or unsatisfactory largely on the basis of their performance on essay exams given at

---

[2] One frequent and interesting result of PC tests is the fact that some individuals go on "witch hunts" of imaginary errors in sentences that are acceptable to most native speakers as well as to linguists. On the Fossil Test, for example, several native speakers of English made changes in the sentence I felt my horse to be on solid ground, a sentence which shows Subject-to-Object raising. Noonan (1985) has noted that raising is by no means a universal phenomenon among languages, and Kellerman (1983) has observed that Dutch EFL students will deem sentences with raised constituents to be less grammatical than sentences without such contituents. Native speakers of English may also have such intuitions about raising when certain types of predicates are involved.

[3] Actually, the Ohio State ESL writing program also divides courses into graduate and undergraduate sections, but those divisions are not relevant to the analysis at hand.

the end of each course. These final exams are also scored holistically, again by two readers with a third reader being consulted in cases of disagreements.

The Fossil Test results to be discussed come from the performance of six classes, two at each of the three levels described above. The results, which are presented in Table 3 in terms of the three course levels, show a clear trend: students in 106 found the PC much more difficult than students in 107 and 108 did. Confirmation of this trend is seen in the F statistic from a one-way analysis of variance ($F = 14.46$, $p < .01$). Nevertheless, there is an obvious discrepancy in the trend: While the difference of means between the 106 and 107 groups is large (and significant at the .05 level on a Tukey multiple comparison test), the difference of means between the 107 and 108 groups is small (and statistically non-significant).

Table 3

Items Correct on PC

	106	107	108
Mean	5.0	9.3	10.1
S.D.	2.3	3.4	4.0
N	21	24	21

One-way ANOVA results: $F = 14.46$, $p < .01$; Maximum possible score = 24

The negligible difference between the 107 and 108 groups might suggest a ceiling effect on the PC. However, results from other groups taking this test suggest that it is indeed possible for non-native speakers to score higher than 10 on the Fossil Test within the 30-minute limit. Consequently, other explanations for the flattening of the linear trend must be sought.

The most likely explanation is that there are only small (if indeed any) differences between the 107 and 108 groups, at least with respect to whatever skills PCs and essay tests both measure. If this explanation is correct, then the PC provides a more accurate indicator of those skills than do the essay tests. The evidence to be presented does support this interpretation, but it also suggests that there are some discrepancies between the two sets of holistic evaluations (for placement tests and for final exams). These discrepancies suggest why the PC and the essay tests give two different pictures of writing ability among 107 and 108 students.

Relevant to these discrepancies are some differences in the student populations in the 106, 107, and 108 levels. All students in the 106 class had been placed there as a result of the holistic score assigned to their placement exam. In contrast, 16 of the 24 107 students had done well enough on the placement exam to be exempted from 106, while 8 other students had previously taken 106 and had done sufficiently well on the final exam to move into 107. These 8 students will be termed "107 move-throughs." Of the 21 students in 108, only four had been placed there as a result of the holistic score on their placement exam, while the others had already taken 107, had passed

the final exam, and had moved into 108. These students will be termed "108 move-throughs." (See Brown, 1981, for a similar analysis.)

Table 4

Perforance by 107 Students on PC
as a Function of Entry into 107

	106M	107P
Top half	2	10
Bottom half	6	6

106M = Previous enrollment in 106
107P = No previous enrollment in 106
Chi-square = 4.69, $p. < .05$

The proportions of 107 and 108 move-throughs within their respective classes are clearly different, and a natural inference is that there is at best only a small difference between students enrolled in 107 as a result of their placement exam scores and students who have moved into 108 on the strength of their performance on the 107 final exam. In other words, it may be the case that there is no significant difference between certain writing skills of 107 and 108 students. Corroboration for this inference comes from a comparison of the 107 students who were move-throughs from 106 and those students enrolled in 107 on the basis of their placement test scores. Table 4 classifies all 24 107 students into the half that scored above the median 107 score on the PC (which was 9.5) and the half that scored below the median. It is clear that a higher proportion of 107 move-throughs are in the bottom half and that a higher proportion of students originally placed into 107 are in the top half. The proportional differences are significant on a chi-square test (4.69, $p < .05$), and the evidence thus suggests that the PC results have to be interpreted in light of how students came to be in 107 and 108.

Such an interpretation is tantamount to saying that the two holistic scoring procedures, for the placement test and for the final exam, do not really give the same results even though the procedures superficially seem the same. A student essay written for a final exam appears more likely to elicit a more favorable evaluation than the same essay would if it were written on a placement test. This discrepancy in evaluations probably has two sources. First, the final exam measures performances of individuals considered to have somewhat comparable writing abilities, whereas the placement exam measures performances of individuals considered to have a wider range of abilities (from 106 to 108). A natural outcome, then, is that the 107 performances on the final will seem more homogeneous. In fact, the grading patterns in the ESL program support this interpretation: very few individuals have stood out either positively or negatively such that few individuals have received either A's or failing grades. Thus most individuals in 107 would be likely to pass, albeit without flying colors,

and to enroll in 108.[4] The second explanation is that the raters of final exams often have less training and experience than do the raters of the placement exams. It is thus possible that evaluations of the final exam are less dependable because of the relative inexperience of some of the graders. There is in fact anecdotal evidence supporting this explanation although space does not permit discussion of that evidence.

Relationships Among Individual PC Items

The final problem to be discussed involves item analysis. The specific data to be considered come from one administration of the Fossil Test at the University of Texas. The internal consistency of the test as measured by KR-20 is moderately high: .80. At the same time, however, the inter-item correlations are generally low and frequently non-significant, even in the case of correlations between items involving the same structural error. For example, only one correlation is significant between three items involving errors of third-person-singular in the present tense (e.g., Such rock often show many fossils...), the three correlations being .07, .09, and .61. It seems likely that a number of factors must be considered in determining why individuals who can detect one error involving the third-person-singular do not detect the others. Differences in the sentence structures within which the errors appear are one possible explanation. However, another factor that probably plays a larger role is the fact that some individuals concentrated their detection efforts on earlier parts of the text while others focused on the last two pages, where two of the three errors were. (The focus of such individuals is evident from the fact that on several individuals' papers many errors were often marked on one page with few errors being marked on other pages.) Aside from these differences of focus, there are quite possibly other types of reading strategies which could strongly influence what one detects, especially in a long text such as the Fossil Test. This is clearly one area where more research on PCs would be useful.

Yet while the inter-item correlations are low, the KR-20 suggests there is considerable internal consistency in the results. The KR-20 estimate is corroborated by the results from 23 multiple regressions in which each test item was the dependent variable and the remaining items the independent variables.[5] The R-Square coefficients obtained from these regressions ranged from a minimum of .50 to .97 with the mean R-Square being .76. Thus the regression equations were normally able to predict about 76 percent of the variance on an item. In other words, performance on any given PC item was generally predicted rather well by collective results from every other item.

---

[4] There have been recent changes in grading practices so that some of the observations made here about the writing program are now purely historical.

[5] One item with a variance of zero was not included in the regression analyses.

References

Arthur, B. 1980. Gauging the boundaries of second language competence: A study of learner judgments. Language Learning 30, 177-195.

Bowen, J. D. 1978. The identification of irrelevant lexical distractions: An editing task. TESL Reporter 12 (1), 1-3, 14-16.

Brown, J.D. 1981. Newly placed students versus continuing students: Comparing proficiency. In J.C. Fisher, M. A. Clarke, and J. Schachter (Eds.), On TESOL '80, building bridges: research and practice in teaching English as a second language. Washington, D.C.: TESOL.

Davies, A. 1975. Two tests of speeded reading. In R. Jones & B. Spolsky (Eds.), Testing language proficiency, (pp. 119-127). Arlington, VA: Center for Applied Linguistics.

Kellerman, E. 1983. Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), Language transfer in language learning, (pp. 112-134). Rowley, MA: Newbury House.

Noonan, M. 1985. Complementation. In T. Shopen (Ed.), Language typology and syntactic description, (pp. 42-140). Cambridge: Cambridge University Press.

Odlin, T. 1985. Passage correction as a measure of writing skills. Paper presented at the the 19th Annual TESOL Convention, New York.

Odlin, T. 1986. Another look at passage correction tests. TESOL Quarterly 20, 123-130.